Exercise 4: Hierarchical Models

Saturday, March 30, 2019 1:17 PM

Hierarchical Models: Data-Analysis Problems

Math tests

1.

We notice that the school with highest mean score (#67) has only 4 students, and the school with the 3rd lowest mean score (#17) has only 7 students. This is due to the fact that computation of mean is total scores divided by number of students, and the smaller that count is, the more the mean is getting affected by extreme score case(s).

2.

We have

$$\theta_i \sim N(\mu, \tau^2\sigma^2)$$

$$y_{ij} \mid \theta_i \sim N(\theta_i, \sigma^2)$$

Denote $\lambda = 1/\sigma^2$, $w = 1/\tau^2$, we have

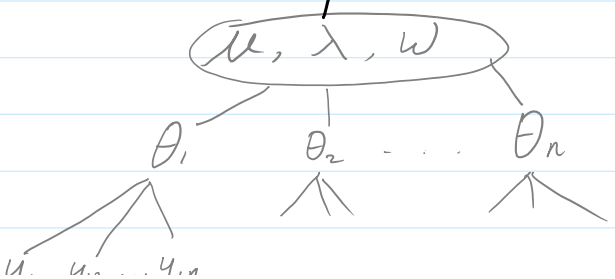
$$f(\theta_i) \propto \exp\left(-\frac{1}{2} w\lambda (\theta_i - \mu)^2\right)$$

$$f(y_{ij} \mid \theta_i) \propto \exp\left(-\frac{1}{2}\lambda (y_{ij} - \theta_i)^2\right)$$

Method 1: Empirical Bayes

We apply a more "frequentist" mindset first, and implement the Empirical Bayes method: compute the marginal likelihood of data y_{ij}'s given hyperparameters μ, λ, w, by integrating out the parameter θ_i's. We then tune the hyperparameters to find the best values in terms of giving the largest marginal likelihood of data, through optimization method like Nelder-Mead.

Note that we call μ, λ, w hyperparameters in the sense that they are the "global" parameters that related to each data point, as opposed to local parameter θ_i which only connects to students from School i:



We derive the marginal likelihood as follows.

$$p(y \mid \mu, \lambda, w) = \int_{\theta \in \mathbb{R}^n} p(y \mid \theta, \mu, \lambda, w) \cdot p(\theta \mid \mu, \lambda, w)\, d\theta$$

(by independence of θ_i's)

$$= \prod_{i=1}^{n} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} p(y_{ij} \mid \theta_i, \lambda) \cdot p(\theta_i \mid \mu, \lambda w)\, d\theta_i$$

in which for each θ_i we have the integration as

$$\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} \lambda^{\frac{1}{2}}\right)^{n_i} \exp\left(-\frac{\lambda}{2} \sum (y_{ij} - \theta_i)^2\right) \cdot \frac{1}{\sqrt{2\pi}} \lambda^{\frac{1}{2}} w^{\frac{1}{2}} \exp\left(-\frac{\lambda w}{2} (\theta_i - \mu)^2\right) d\theta_i$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right)^{n_i} \lambda^{\frac{n_i}{2}} \exp\left(-\frac{\lambda}{2}\left[\sum y_{ij}^2 - 2\sum y_{ij}\theta_i + n_i \cdot \theta_i^2\right]\right) \cdot \frac{1}{\sqrt{2\pi}} \lambda^{\frac{1}{2}} w^{\frac{1}{2}} \exp\left(-\frac{\lambda w}{2}(\theta_i^2 - 2\mu\theta_i + \mu^2)\right)$$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right)^{n_i} \lambda^{\frac{n_i+1}{2}} w^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\lambda \sum y_{ij}^2 + \lambda w \mu^2)\right) \cdot$$

$$\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\lambda(n_i+w)}}{\sqrt{\lambda(n_i+w)}} \exp\left\{-\frac{1}{2}\left[(\lambda n_i + \lambda w)\left\{\theta_i^2 - 2\frac{\lambda \sum y_{ij} + \lambda w \mu}{\lambda n_i + \lambda w}\theta_i + \left(\frac{\sum y_{ij} + \lambda w \mu}{n_i + w}\right)^2 - \left(\frac{\sum y_{ij} + \lambda w \mu}{n_i + w}\right)^2\right\}\right]\right\} d\theta_i$$

$$\propto \lambda^{\frac{n_i+1}{2}} w^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\lambda \sum y_{ij}^2 + \lambda w \mu^2 - \frac{\lambda(\sum y_{ij} + w\mu)^2}{n_i + w}\right)\right) \left(\lambda(n_i+w)\right)^{-\frac{1}{2}}$$

$$\cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} (\lambda(n_i+w))^{\frac{1}{2}} \exp\left(-\frac{1}{2}\left[(\lambda n_i + w)\theta_i^2 - 2(\lambda \sum y_{ij} + \lambda w \mu)\theta_i + \frac{\lambda(\sum y_{ij} + w\mu)^2}{n_i + w}\right]\right) d\theta_i$$

pdf of $\mathcal{N}\left(\frac{\sum y_{ij} + w\mu}{n_i + w},\ (\lambda(n_i+w))^{-1}\right)$

$$= \lambda^{\frac{n_i}{2}} \left(\frac{w}{n_i + w}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\lambda \cdot \sum y_{ij}^2 + \lambda w \mu^2 - \frac{\lambda(\sum y_{ij} + w\mu)^2}{n_i + w}\right]\right\}$$

This computation might lead to data overflow because we observe terms like $\sum y_{ij}^2$ in the exponent, thus we take the log of it to get the marginal log-likelihood of data w.r.t. School i:

$$\frac{n_i}{2}\log\lambda + \frac{1}{2}\log\frac{w}{n_i+w} - \frac{1}{2}\left[\lambda \cdot \sum y_{ij}^2 + \lambda w \mu^2 - \frac{\lambda(\sum y_{ij} + w\mu)^2}{n_i + w}\right]$$

and the marginal log-likelihood is thus given by

$$\log p(y \mid \mu, \lambda, w) = \sum_{i=1}^{n} \left\{ \frac{n_i}{2} \log \lambda + \frac{1}{2} \log \frac{w}{n_i + w} - \frac{1}{2} \left[\lambda \cdot \sum y_{ij}^2 + \lambda w \mu^2 - \frac{\lambda (\sum y_{ij} + w\mu)^2}{n_i + w} \right] \right\} + \text{constant}.$$

which is our objective function to be maximized.

We use Scipy's implemenration of Nelder-Mead method to minimize the negative log-likelihood, which gives the optimal values of μ, λ, w instantly, which are extremely close to the Markov chain estimation from Gibbs sampler. Detailed implementation is in SDS_383D_Exercise_4.ipynb for both methods.

Method 2: Gibbs Sampler

We can also impose appropriate prior distributions on parameters μ, λ, w, and thus derive the posterior conditionals from which we can sample each parameter iteratively: i.e. the Gibbs Sampler framework.

We set

$$\lambda \sim \text{Gamma}(a, b)$$
$$w \sim \text{Gamma}(c, d)$$
$$\mu \sim N(\nu, \phi^{-1})$$

Denote n_i to be the number of students in school i, we have the joint posterior of $\theta_{1:n}, \lambda, \mu, w$ as

$$f(\theta_{1:n}, \mu, \lambda, w \mid y_{1:n, 1:n_i}) \propto f(y_{1:n, 1:n_i} \mid \theta_{1:n}) \, f(\theta_{1:n}) \, f(\mu) \, f(\lambda) \, f(w)$$

$$\propto \prod_{i=1}^{n} \prod_{j=1}^{n_i} f(y_{ij} \mid \theta_i) \cdot \prod_{i=1}^{n} f(\theta_i) \cdot f(\mu) \, f(\lambda) \, f(w)$$

$$= \prod_{i=1}^{n} \left(\prod_{j} \exp\left(-\frac{1}{2} \lambda (y_{ij} - \theta_i)^2\right) \right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2} w \lambda (\theta_i - \mu)^2\right)$$
$$\cdot \exp\left(-\frac{\phi}{2} (\mu - \nu)^2\right) \lambda^{a-1} e^{-b\lambda} \, w^{c-1} e^{-dw}$$

We then go ahead and pick the terms containing each parameter to derive their posterior conditionals:

θ_i: the posterior density for each θ_i has the form

$$f(\theta_i|\cdots) \propto \exp\left\{-\frac{1}{2}\left[(\lambda n_i + \lambda w)\theta_i^2 - 2\lambda(\Sigma y_{ij} + w\mu)\theta_i\right]\right\}$$

thus $\theta_i|\cdots \sim N\left(\frac{\Sigma y_{ij} + w\mu}{n_i + w},\ \lambda^{-1}(n_i + w)^{-1}\right)$

μ: $f(\mu|\cdots) \propto \exp\left\{-\frac{w\lambda}{2}\sum_{i=1}^{n}(\theta_i - \mu)^2\right\} \cdot \exp\left(-\frac{\phi}{2}(\mu - v)^2\right)$

$$\propto \exp\left\{-\frac{1}{2}\left[(w\lambda n + \phi)\mu^2 - 2\left(w\lambda(\textstyle\sum_i \theta_i) + \phi v\right)\mu\right]\right\}$$

thus $\mu|\cdots \sim N\left(\frac{w\lambda(\sum_{i=1}^{n}\theta_i) + \phi v}{w\lambda n + \phi},\ (w\lambda n + \phi)^{-1}\right)$

λ: $f(\lambda|\cdots) \propto \lambda^{\left(\sum_{i=1}^{n} m_i + n\right)/2 + a - 1} \exp\left\{-\left[\frac{1}{2}\left(\sum_i\sum_j (y_{ij} - \theta_i)^2 + w\sum_i(\theta_i - \mu)^2\right) + b\right]\lambda\right\}$

(m_i: # of students in School i)

thus $\lambda|\cdots \sim \text{Gamma}\left(\frac{\sum_{i=1}^{n} m_i + n}{2} + a,\ \frac{\sum_i\sum_j (y_{ij} - \theta_i)^2 + w\sum_i(\theta_i - \mu)^2}{2} + b\right)$

w: $f(w|\cdots) \propto w^{\frac{n}{2}} \cdot \exp\left(-\left(\frac{\lambda\sum_i(\theta_i - \mu)^2}{2} + d\right)w\right) w^{c-1}$

$$w|\cdots \sim \text{Gamma}\left(\frac{n}{2} + c,\ \frac{\lambda\sum_i(\theta_i - \mu)^2}{2} + d\right)$$

3.

Check ipython notebook for code and plot.